nature research

Corresponding author(s): Christina Curtis

Last updated by author(s): Apr 7, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data sources are reported in the maintext and Methods. Briefly, we performed a comprehensive review of the literature by surveying PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) to identify studies in which whole-exome sequencing (WES) was available for matched normal tissues, primary tumors (P) and metastases (M). We focused on colorectal, lung and breast cancers given the availability of large cohorts for these three common cancer types. In total, the raw sequencing reads data for 586 tumor samples from 181 patients in 13 published studies were accessed and retrieved (Supplementary Table 1). We also generated multi-region sequencing (MRS) data for two colorectal cancer patients (mCRCTB1 and mCRCTB7) with liver metastases for who de-identified excess tissue was collected as part of routine care, using the Agilent SureSelect Human All Exon kit for sequencing on the Illumina HiSeq 2500.
Data analysis	Bioinformatic tools and version numbers are reported in the Methods. All statistical analyses were conducted in the R environment (v3.5.0). Custom codes for modeling tumor evolution and estimating the timing of metastasis will be published with the manuscript (https://github.com/cancersysbio/pan-metastasis). In particular, the bioinformatic and computational tools used in this paper are: BWA (v.0.7.10), Picard Tools (v.1.111), GATK (3.4.0), MuTect (v.1.1.7) ,Varscan (v.2.3.9), DToxoG (1.14.4.0), SAMtools (v.1.2), ANNOVAR (v.20150617), TitanCNA (v.1.5.7), HMMcopy (v.0.99.0), CHAT (v 1.0), dndscv, FigTree (v1.4.4), PolyPhen-2(http://genetics.bwh.harvard.edu/pph2/), CHASMplus (http://fathmm.biocompute.org.uk/fathmm-xf/), CHASMplus (https://karchinlab.github.io/CHASMplus/), Metascape (http://metascape.org), MuSiCa (http://bioinfo.ciberehd.org:3838/MuSiCa/) and PHYLIP (http://www.trex.uqam.ca/index.php?action=phylip&apP=dnapars).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Sequencing data have been deposited at the European Genome Phenome Archive (EGA) under the accession number EGAS00001003573.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	In total, 457 tumor samples from 136 metastatic cancer patients including 39 colorectal cancers (181 tumor samples), 30 lung cancers (74 tumor samples) and 67 breast cancers (202 tumor samples) were included in this study after quality control assessment. Further information on these samples is described in Supplementary Tables 1-2.
Data exclusions	Cases (n=5) with none of or very few (<10) trunk SSNVs/indels were excluded as which implies independent (non-clonal) origin for the primary tumor and metastasis. Furthermore, patients (n=42) with a diffusely distributed cluster for truncal SSNVs/indels were also excluded since this is likely caused by low tumor purity or low sequencing quality. These exclusion criteria were not pre-established but the they are required for our research purpose. After these filtering steps, 457 tumor samples from 136 metastatic cancer patients including 39 colorectal cancers (181 tumor samples), 30 lung cancers (74 tumor samples) and 67 breast cancers (202 tumor samples) were retained for downstream analysis in this study.
Replication	The number of replicates in simulation studies are reported in the figure legends and Methods. In particular, 500 virtual P/M pairs were simulated under each of the monoclonal seeding and polyclonal seeding scenarios. 1000 virtual primary tumors were simulated to estimate the parameter alpha in Eq.(1).
Randomization	This study does not involve randomization as it is based on available sequencing data for paired primaries and metastases.
Blinding	This study does not involve group allocation as it is based on available sequencing data for paired primaries and metastases.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging